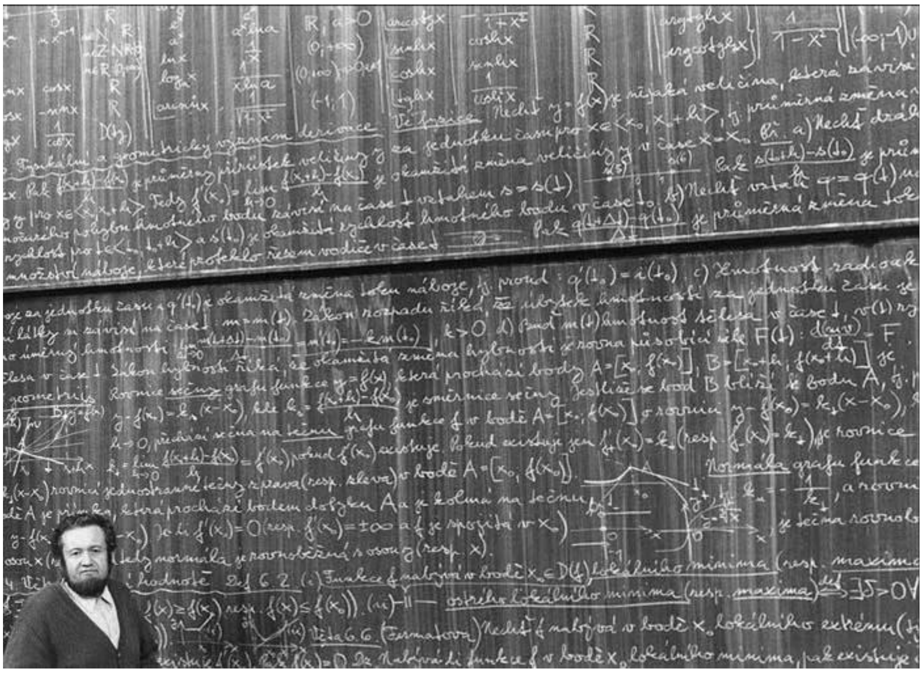
GREATER DANDENONG
City of Opportunity

Elementary Statistics for Social Planners



Hayden Brown
Social Policy Unit
2013

CONTENTS

Elementary Statistics

To type equations: Word > Insert > Objects > Select 'Microsoft Equation Builder 3.0

I: Tables

Dependent variable in a table – in the rows
In a title – first part of heading (e.g.: income by education)
In a chart – the Y axis
In a linear regression: y

In the table, present the percentages across the categories of dependent variable (that is, percentages adding to 100 down the columns). That way, one may easily discern how the distribution of cases across the dependent variable varies for each category of independent variable.

Eg: Purchasing patterns by Educational Level

Educational level (the Independent variable)

Dependent var	**Primary**	**Secondary**	**Tertiary**
Blue collar	70%	62%	44%
White collar	30%	38%	56%
Total	100%	100%	100%
Number	89	100	77

Note

- Actual number of cases should be given at the bottom of each column.
- Add footnotes with any details of category definition and any anomalies such as missing data or rows, or columns not adding to 100

II: Types of social data

- Nominal measures – different categories
- Ordinal measures – ordered categories
- Interval – ordered categories of equal width

More complex or higher order data is preferable, as it contains more information.
Not always clear which type of data one is dealing with. For instance, of income:
* May merely represent a lifestyle = nominal
* Differentiates respondents by social prestige or each extra dollar means less, as a persons wealth increase = ordinal
* Each dollar is as large as the next in monetary terms ⇨ interval

Arguably, many social variables much as income, age and education, appear to be interval variables, but are more ordinal because the social significance of each unit – dollar, year , year of education – is not the same as every other.

Similarly, Likert and other rating scales may be ordinal because respondents do not appear to consider each gradation as equal: e.g.: they tend to treat extreme negative responses as larger increments than extreme positive responses (which are easier to give because they seem more congenial)
One may always treat higher-order information as lower-order – for instance, interval data as ordinal or nominal – because higher-order information contains more data than lower-order. One however, treat lower-order data as higher-order, as it does not contain enough information.

Characteristics of Social Measures Measures should be:

- exhaustive and mutually-exclusive, where used as categories of a variable
- valid – meaning what the respondent thinks they mean
- reliable – each case is coded the same way by a given coder (intra-coder reliability) and by other coders (intercoder reliability); and otherwise measured in a consistent way

Measures of cause

Often, in everyday life, a particular decision or action is caused or contributed to, by a range of factors. And frequently, no single factor alone was either sufficient or necessary for that outcome. Some researchers point to several types of causes:

- What *caused* respondents to do…
- What *purpose* the respondents had in doing…
- What *enabled* the respondent to do…

So to understand why someone does, believes or favours something, either researcher or respondent should specify the perspective or frame of reference used in answering the question: 'Why?'

Contents of these notes

	Nominal	Ordinal	Interval
Central Tendency	Mode	Median	Mean
Dispersion	Variation ratio Index of qualitative variation Index of diversity	Interquartile range Inter-decile range	Standard deviation
Association between two variables	Proportional reduction in error & Landba Chi Square	Gamma Spearman's rho	Pearson's r
Significance & strength of association	Chi Square	Significance of Gamma Significance of Spearman's	Significance of r & Linear Regression

III: Measures of Central Tendency

Mode – Nominal data: The category with the highest frequency (⇨ your best guess).

Median – Ordinal data: The middle value or 50th Percentile

For grouped data: $$Median = L + \frac{(N/2 - F)}{fm} xh$$

where L = lower limit of the interval
N = total cases
F = total cases below interval
Fm = interval frequency (ie: cases in interval)
H = width of interval

Arithmetic Mean = sum of cases / number of cases $\bar{x} = \frac{\sum x}{n}$ ($\bar{x}$ for a sample; μ for a population)

Note:

- If data is skewed it will pull median in the direction of skew, in which case, a median may be preferable to a mean, or in extreme skew, a mode.
- Extreme scores affect means more than medians. So extreme scores make the median preferable
- If data is rectangular or multi-modal, mean makes little sense; better to provide a whole distribution.

IV: Measures of Dispersion

Nominal Data

Variation ratio: VR = 1 – modal frequency/n ie: proportion of cases that are non-modal

- But VR does not consider the distribution of cases among the non-modal categories
- Only good for comparing dispersion where you have the same number of categories in each table

Index of Qualitative Variation: IQV = total number of observed differences/ max. no of possible differences
The difference being counted in the top and bottom of the equation are the number of possible pairs of cases, where each case is in a different category.

E.g.: if of 15 people, 9 owned their homes, 3 were renting and 3 paying a mortgage, the there would be (9x3) + (9x3) + (3x3) = 63 ways in which pairs of cases could be made with each having voted for a different party. This is the number of observed differences: the numerator of the formula.
However, the maximum number of such pairs would arise be created if the distribution was rectangular – that is, equal numbers in each category; in this instance 6, where there would be (5x5) + (5x5) + (5x5) = 75 different pairs. This is the number of possible differences: the denominator of the equation.
⇨ IQV = 63/75 = 0.84
Note:

- IQV varies from 0 to 1 and higher values equal higher variations
- Only comparable for distributions with the same number of categories – just like the VR Ratio

Computational formula: $IQV = \dfrac{k(n^2 - \sum_{i=1}^{k} f_i^2)}{n^2 (k-1)}$

where k = total categories
n = number of cases
f_i = frequency of i^{th} category

For the example above, IQV = $IQV = \dfrac{3(n^2 - (5^2 + 3^2 + 3^2))}{15^2 x(3-1))}$ = 0.84

Index of Diversity

For measuring the dispersion of data in nominal categories, one may determine the average of the probability that any two cases randomly selected from a sample of nominal data would be from different categories.

For each category, the proportion of all cases in that category is calculated as the number of cases in the category (C) divided by N or C/N. Since this probability – that a randomly selected case would belong to this category - applies to all members of that category, C/N is multiplied by C, giving the sum of the probabilities for each case in that category. This is repeated for all other categories, then the sum is divided by N, to give the average of the probability that two cases from the sample selected at random, would be from the *same* category. This value is then deducted from 1, to give the probability that two cases selected at random would be from a *different* category. The formula is:

$= 1 - \dfrac{\sum_{i=1}^{c} f_i^2}{N^2}$ Where c = number of categories, f_i = frequency of the i^{th} category, N = sample size.

Ordinal Data

Inter-decile range: covers 80% of observations = d9 – d1 (d9 = 9^{th} decile, below which 90% of cases lie)
Interquartile range: covers 50% of observations = q3 – q1 (q3 = 3^{rd} quartile, below which 75% of cases fall)

E.g.: Literacy

Rank	Literacy Level	% respondents	
1	Level 1 (lowest)	9	
2	Level 2	29	- d1
3	Level 3	41	
4	Level 4	38	
5	Level 5 (highest)	17	- d9

Here the inter-decile range ⇨ d9-d1 = 5-2 = 3

Interval Data

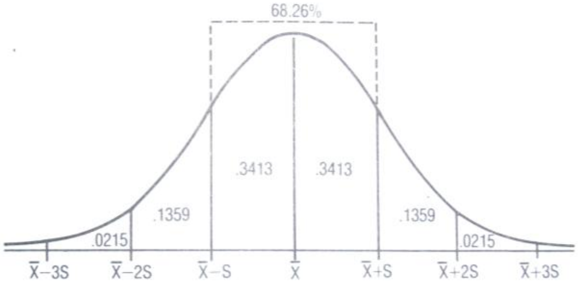
Standard deviation $s = \sqrt{\sum_{i=1}^{n} \frac{(x_i - \bar{x})^2}{n}}$

where n = number of cases
x_i = ith value
$\bar{x}$ = mean of values

No matter how the original cases are distributed – **75%** of observations fall within the range $\bar{x} \pm 2s$
89% of observations fall within the range $\bar{x} \pm 3s$

In a Normal distribution though: **68%** of cases fall within range $\bar{x} \pm 1s$
95% of cases fall within range $\bar{x} \pm 2s$

For each Z score, the area under the curve between the mean and that point can be looked up in a table. The area between the mean and the s score of 1, would cover 34% of the area under a normal curve, while the area between a z score of -1 and 1 would encompass 68% of cases.



Z, or standard score, is the distance of a value from the mean, in standard deviation units. $Z = \frac{x_i - \bar{x}}{s}$

V: Random Samples

Random – a random selection of units from a defined population.
In a survey, this means you have to randomly select individuals and receive a response from them.

Symbols:

x	any variable, in univariate distributions
X_i	any x score
X_1	first x score in the sample
X_n	last x score in the sample
Y	dependent variable in bivariate distributions

	Mean	SD	Variance	Sample size
Sample	$\bar{x}$	s	s^2	n
Population	μ	σ	σ^2	N

Measure for a sample is called a statistic: e.g.: $\bar{x}$, s
Measure for the population is called a parameter: e.g.: μ, σ

Random Sampling

Suppose one took a sample from a population, which had a mean of a mean of $\bar{x}$ and a standard deviation of s. If repeated samples of size n had been drawn from a population with a mean μ and standard deviation of σ, then you would end up with a distribution of these sample means. This distribution would have:

- A mean of μ
- A standard deviation (called standard error, or se_x): $se_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ But if n>30, $\sigma = s$, so $se_{\bar{x}} = \frac{s}{\sqrt{n}}$

So, the single sample taken would have a mean of $\bar{x}$, which is a member of this imaginary distribution of sample means. That distribution is a normal distribution, with a mean of μ, as mentioned above.

Therefore, there is a 95% chance that $\bar{x} = \mu \pm 1.96 \text{ x } se_x = \mu \pm 1.96 \text{ x } \frac{s}{\sqrt{n}}$

Rearranging the equation ⇨ 95% chance that $\mu = \bar{x} \pm 1.96 \text{ x } se_x = \bar{x} \pm 1.96 \text{ x } \frac{s}{\sqrt{n}}$

As the sample size (n) increases, the standard deviation $\frac{s}{\sqrt{n}}$ declines, making the estimate of the population mean more precise. So sample size, not the population, determines the precision of the population estimate.

E.g.: a 5% sample of 29,000 rate payers, or numbering 1,150 ratepayers, were surveyed for their support for a multi-storey development in their neighbourhood. The mean response in a five-category scale from ‘very supporting’ to ‘very opposed’ was 3.9 and s= 2.1

⇨ se = 2.1/√1,150 = 0.076

$\bar{x} = 3.9$

⇨ there is a 95% chance that $\mu = 3.9 \pm 1.96 \text{ x } 0.076$
$= 3.9 \pm 0.15$
$= 3.75 \text{ to } 4.05$

Sample size required to estimate μ with a degree of precision

Recall: there was a 95% chance that $\mu = \bar{x} \pm 1.96 \text{ x } \frac{s}{\sqrt{n}}$

* 1.96 x $\frac{s}{\sqrt{n}}$ may be considered the margin or error, which we will label as E

* 1.96 is the Z score corresponding to the confidence level required. We will label it Z

* n = sample size required to give a result within this margin of error with a certain confidence level

So $E = 1.96 \text{ x } \frac{s}{\sqrt{n}}$ ⇨ $E = Z \text{ x } \frac{s}{\sqrt{n}}$ ⇨ $n = \left(\frac{Z * s}{E}\right)^2$

Eg: To estimate weekly income within ± \$60 at a confidence level of 95%, when s = \$290.

Required sample size: $n_s = (\frac{\sigma x Z}{E})^2$ ⇨ $n_s = (\frac{290x1.96}{60})^2 = 89$

Standard Error of a Percentage

For a percentage value (pc) s = $\sqrt{pc * (100 - pc)}$ Substituting this into $se_{\bar{x}} = \frac{s}{\sqrt{n}}$ …

⇨ $se_{pc} = \sqrt{\frac{pc - (100 - pc)}{n}}$ or the standard error of a distribution of sample percentages

Now that the se of the distribution of sample percentages has been estimated, one may be 95% sure that the population percentage lies win the range: sample percentage ± 1.96 se_{pc}

Size of sample to estimate population percentage with a degree of precision:

For a percentage value (pc) s = $\sqrt{pc * (100 - pc)}$

Since $n_s = \left(\frac{Z * s}{E}\right)^2$ Then $n_s = \frac{Z^2 * pc(100 - pc)}{E^2}$

Where: pc = population percentage based on a preliminary study
Z = se corresponding to a specified confidence level e.g.: 1.96 for 95% confidence
E = maximum percent that the estimate can be from the true population percentage
N_s = size of sample required

VI: Association Between Variables

NOMINAL VARIABLES

Proportional Reduction of Error

If you had data about religion, such as this:

India	Thailand	England	Total
48	32	16	96

And you had to give a prediction about birthplace, you would guess ‘India. You would be wrong though, 48 /96 times, or half the time. However, if the table was cross tabulated with skin colour…

	India	Thailand	England
Hindu	40	3	4
Buddhist	1	23	3
Christian	7	6	9

…and you could use skin colour to help you guess a person’s religion, you would guess Christian for English-born people, Buddhist for Thai and Hindu for Indian – and you would be wrong 8+9+7 =24 times out of 96.

$$PRE = \frac{originalerror - finalerror}{originalerror}$$ which in this case, = (48-24)/48 = 24/48 = 0.5

Landba

Calculated in an identical faction to PRE, above, though with a computational formula. $\lambda_a = \frac{\sum f_i - F_d}{n - F_d}$

Where f_i = modal frequency within each category of independent variable
Fd = modal frequency among totals of the dependent category variables
n = number of cases

If 1, then all error is removed by knowing the independent variable. If 0, no error is removed.

Eg: if for each category of the independent variable the greater number of cases is in the same category of dependent variable, then there is not reduction of effort, so lambda will = 0. In other words, if for each of the dependent variables, the modal category of dependent variable is the same, Lambda = 0, so Lambda will not help.

		Unemployed	Employed
Agree	1	55	72
	2	18	54
Disagree	3	6	1

Chi Square

For calculating the statistical significance of the association between two nominal variables

Compare frequency distribution in a cross tab, with the distribution expected from a sample drawn from a no-relationship population. That is where for each category of the independent variable (columns, by convention) the distribution of cases across the dependent variable (rows) is the same.

Actual

	Education Level			
Employment	**Low**	**Medium**	**High**	**Total**
Jobless	23	15	6	44
Employed	10	14	29	53
Total	33	29	35	97

Expected

	Education Level			
Employment	**Low**	**Medium**	**High**	**Total**
Jobless	15	13	16	44
Employed	18	16	19	53
Total	33	29	35	97

Procedure

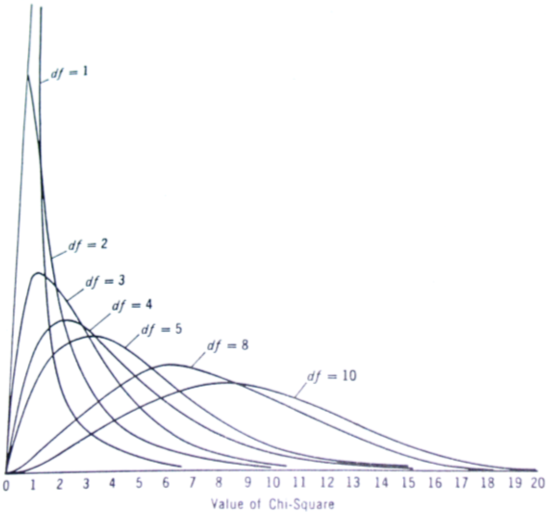
1. Calculate expected values or each value in table

The ratio between E and its column total should be equal to the ratio between the corresponding row total and the total cases. Therefore E/column total = row total / N ⇨ E = (column total x row total) / N

2. Calculate $\chi^2 = \sum \frac{(O-E)^2}{E}$ where O = observed value; E = expected value

Ie: sum of the ratios between the square of difference between each value and its expected frequency, and the expected frequency.

3. Calculate the degrees of freedom: df = (columns – 1) x (rows – 1)



4. The χ^2 value lies on a one-tailed distribution of χ^2 values for a non-relationship situation, for that particular number of df. If χ^2 lies in the tail, near the far right of the distribution, where say 5% of the χ^2 are situated, then that value is unlikely to have occurred by chance in a non-relationship situation. Therefore it may be concluded that there is, in fact, a significant association between the two nominal variables.

So, look up the threshold value of χ^2 (for it to represent a statistically significant value) for that number of df and for the confidence level required (typically α = 5% (0.05) or 1% (0.01)). This confidence level represents the possibility of a type 1 error, where you conclude that there is a relationship and reject H_o (the null hypothesis), when in fact there is no relationship.

Eg: If χ^2 for a 3 x 4 column table = 13.5

- $\chi^2 = 13.5$
- df = 2 x 3 = 6
- $\alpha = 5\%$

⇨ critical value = 12.59 So a χ^2 of 13.5 indicates a 95% probability that H_o (null hypothesis) is False (i.e.: χ^2 of 13.5 will occur 5% of the time when the null Hypothesis is true)

In Excel, Chidist(χ^2, df) gives the proportion of the χ^2 distribution for that df which lies to the right of that value. If that value is <0.05, you may accept that there is a relationship.

ORDINAL VARIABLES

Gamma: [**for grouped ordinal data of <7 or so categories]** A table of data is set up, so that the order of the categories of both variables increases as you proceed to the right and downward.

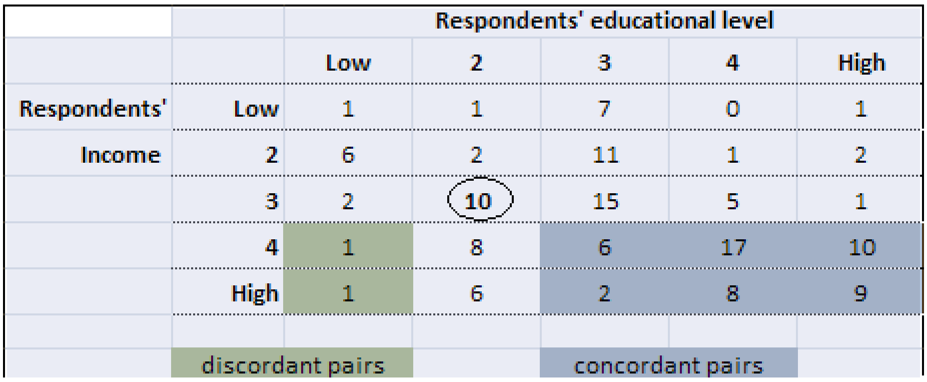
First, for each cell, the number of cases (its value) is multiplied by all of the values of the cells below and to its right. This gives a count of all *concordant* pairs – that is, all pairs of cases that are consistent with the idea that an increase in one variable is associated with an increase in the other - a positive relationship between the variables.

Second, for each cell, its value is multiplied by the values in each cell below and to its left, representing *discordant* pairs – where a high value on one variable is associated with a lower value on another – a negative relationship.

Once these concordant and discordant pairs are counted, G is calculated, so: $G = \dfrac{f_a - f_i}{f_a + f_i}$

Where f_a = number of concordant pairs
f_i = number of discordant pairs

In the table below, the value of the cell circled is multiplied by the sum of the values below and to its right to give a total of the *concordant* pairs for that cell = 10 x (6+17+10+2+8+9) = 520
And the circled cell is multiplied by the sum of cells below and to its left = 10 x (1+1) = 20 = *discordant* pairs.



		Respondents' educational level				
		Low	2	3	4	High
Respondents'	Low	1	1	7	0	1
Income	2	6	2	11	1	2
	3	2	10	15	5	1
	4	1	8	6	17	10
	High	1	6	2	8	9
	discordant pairs			concordant pairs		

The same procedure is adopted for each cell, to give a total of all concordant and discordant pairs.
Clearly if concordant pairs = 0, then G = -1
If discordant pairs = 0 then G = 1

Note

- Only pairs of people with different scores on both variables are counted
- G does not work well if the data does not centre on one or other diagonals. E.g.: if the data were concentrated on the edge of the table
- One may not compare G for tables with different number of categories in either variable.

Significance of Gamma

Calculate Z score for gamma: $Z = G\sqrt{\dfrac{f_a + f_i}{N(1-G^2)}}$

Then, since the critical value for Z ate the 0.05 level is ± 1.96, then a value above this would be significant

Spearman's p (rho) **[for association between ordinal & interval data]** – $\rho = 1 - \dfrac{6\sum D^2}{n(n^2-1)}$

Where: D = difference in number of ranks between an x score and corresponding y score
n = number of cases

To calculate, each score would be ranked on each of the two variables. If several scores were the same, they would be assigned the average of the ranks they would have held if each were different.
Eg: if 3 scores held the rank of 2 and they would have equalled ranks 4, 5 and 6, then they would each be given a rank of 5.
So, for each case, you have a rank on x and on y.
Then, for each case, D is calculated as the x rank minus the y rank. Then it is squared to give D^2
The D^2 is calculated and summed and the p calculated from the formula

Eg	**Rank on x**	**Rank on y**	**D**	**D^2**
	12	7	5	25

Note

- If variables are strongly associated, their ranks will be similar, so D' and D2 will be low and p high.
- If values are skewed, or if there are many ties, p does not work well.
- G, r and p range from -1 to 1 and do not work well with skew

Significance of p: The value of p can be compared with a distribution of p values from a no-relationship population, to determine its significance.

So if either variable is nominal ⇨ use lambda
Ordinal ⇨ use G or p
If both variables are Interval ⇨ use r

INTERVAL VARIABLES

Pearson's r, or Pearson's product moment correlation coefficient may be though of as the proportional reduction in error that would be achieved if y values were estimated by using the line of best fit rather than the average value of y. It may be expressed as:

$r^2 = \frac{originalerror - currenterror}{originalerror} = \frac{\sum(y-\bar{y})^2 - \sum(y-\hat{y})^2}{\sum(y-\bar{y})^2}$ (r is the square root of this)

This may be adjusted to give the formula: $r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i-\bar{x})^2 \sum(y_i - \bar{y})^2}}$

Considering the numerator, values in the upper right of the chart will be the product of two substantial positive numbers, whereas values in the lower left will be a product of two substantial negative values. Either way, the numerator will be positive⇨ a positive correlation (r)
On the other hand, values located in the upper left or lower right hand side of the table will be negative (ie: upper left – $x_i - \bar{x}$ will be negative, and $y_i - \bar{y}$ positive)
Note

- Because the numerator is divided by the difference of all the values from the mean of x and y, r is not affected by how wide a span of values the x and y values are spread across.
- If values are skewed on either variable, it depresses the value of r

Significance of a correlation coefficient

The significance is tested by using the distribution of t: $t = r\sqrt{\frac{N-2}{1-r^2}}$

The number of degrees of freedom = N-2
Once the result for t is obtained, look up the value in a table, for the appropriate degrees of freedom, and see if the result exceeds the value required for significance at the 5% or 1% level.

Eg: where r=0.50 and N=20 $t = .50\sqrt{\frac{20-2}{1-0.50^2}}$ =2.45

In a two-tailed table of t, for degrees of freedom = 18, a value of 2.1 is required for significance at the 5% level and 2.88 at the 1% level. So this result is significant at the 5% level. In other words, such a result would only have been obtained 5% of the time from a no-relationship population.

Correlation tells you the extent to which change in one variable is matched by changes in the other. It does not tell you the actual *extent* to which one variable will change when the other variable changes. For example, if one variable changes very slightly when the other changes markedly, but does so consistently, the correlation may be still be high.

Linear Regression

Allows one to determine the extent to which a change in one interval variable is matched by change in the other. The steps described below produce a calculation of a line of best fit for points on a correlation. This line of best fit takes the form: y = bx + a

* To calculate b: b = S_{xy}/S_{xx} Where $S_{xy} = \sum xy - \frac{\sum x \sum y}{n}$

$$S_{xx} = \sum x^2 - \frac{(\sum x)^2}{n}$$

* Once **b** is known, rather than substituting it into just any point in the scatter plot to determine **a** (since many points do not even sit on the line of best fit) a point which does reside on the best fit line is used. This is the point: ($\bar{x}$, $\bar{y}$) So b is substituted into $\bar{y}$ = b $\bar{x}$ + a, to get a.

Eg: For these values

x	1	2	4	5	7	8
y	1	3	3	5	3	5

$\Sigma x = 27$ $\Sigma y = 20$ $\Sigma xy = 105$ $\Sigma x^2 = 159$

⇨ $S_{xy} = 105 – (27 \text{ x } 20)/6 = 15$
⇨ $S_{xx} = 159 – (27 \text{ x } 27)/6 = 37.5$

⇨ $b = S_{xy}/S_{xx} = 15/37.5 = 0.4$

⇨ $a = \bar{y} – b\bar{x} = 3.3 – 0.4 \text{ x } 4.5 = 1.53$

So line of best fit is: y = bx + a, which in this case is: $y = 0.4x + 1.53$

In Excel, Linest(yrange, xrange) calculates b

√ Σ χ σ μ $\bar{x}$ $\bar{y}$